

Web Appendix

1.1 Case definition

A suspected case is defined as patients meeting ≥ 1 epidemiological criteria and ≥ 2 clinical criteria outlined below.

Epidemiological criteria

- (1) Travel or residence history in Wuhan or nearby cities during 14 days before symptom onset;
- (2) Contact history with a PCR-positive COVID-19 patient during 14 days before symptom onset;
- (3) Contact history during 14 days before symptom onset with patients who had fever or respiratory symptoms and came from Wuhan or communities with reported COVID-19 cases; and
- (4) Related to a cluster of COVID-19.

Clinical criteria

- (1) Fever and/or respiratory symptoms
- (2) Radiographic characteristics of pneumonia, such as multiple ground-glass shadows, infiltrative shadows and consolidation in both lungs;
- (3) normal or lower leukocyte counts, or lower lymphocyte counts at acute phased of the disease.

A confirmed case is defined as a suspected patient with positive detection of 2019-nCoV nucleic acid by real-time RT-PCR or viral genes that are highly homologous to 2019-nCoV by sequencing using respiratory or blood specimens.

1.2 Calculating the effective reproductive number

In this analysis, all non-primary cases are considered as secondary cases. We first partition all cases into four categories: imported primary, local primary, imported secondary and local secondary. We then use a simple moving average approach to estimate the effective reproductive number R_t for the period of Jan. 18-30, 2020. Calculation is limited to this period because secondary cases outside this period are sparse and uncertainty will be too large. For each day within the period, the general estimate for R_t is given by:

$$\hat{R}_t = \frac{N_{sec}(t-2, t+2)}{N_{pri}(t-2, t+2) + \tilde{N}_{sec}(t-2, t+2)},$$

where $N_{pri}(t_1, t_2)$ is the total numbers of primary cases in all clusters whose onset dates were within the time window from day t_1 to day t_2 , and $N_{sec}(t_1, t_2)$ is the total numbers of secondary cases in the clusters of these primary cases. $\tilde{N}_{sec}(t_1, t_2)$ is the number secondary cases in all clusters whose onset dates were within the time window from day t_1 to day t_2 , and we assume all secondary cases infected 0 other people. This assumption is not realistic as some of these secondary cases may be tertiary cases, but it does not affect much the estimation of R_t because potential tertiary cases were attributed to primary cases in $(t-2, t+2)$ or nearby windows. The calculation of $N_{pri}(t_1, t_2)$, $N_{sec}(t_1, t_2)$ and $\tilde{N}_{sec}(t_1, t_2)$ differs by how we allocate imported secondary cases and local primary cases to the numerator or the denominator. Analogous to $N_{pri}(t_1, t_2)$, $N_{sec}(t_1, t_2)$ and $\tilde{N}_{sec}(t_1, t_2)$, we define $N_{pri}^{imp}(t_1, t_2)$, $N_{sec}^{imp}(t_1, t_2)$ and $\tilde{N}_{sec}^{imp}(t_1, t_2)$ for imported primary and secondary cases, as well as $N_{pri}^{loc}(t_1, t_2)$, $N_{sec}^{loc}(t_1, t_2)$ and $\tilde{N}_{sec}^{loc}(t_1, t_2)$ for local primary and secondary cases. Unlike $N_{pri}(t_1, t_2)$, $N_{sec}(t_1, t_2)$, and $\tilde{N}_{sec}(t_1, t_2)$, these category-specific numbers are observed and fixed. R_t is calculated in the following three scenarios:

- (1) All imported cases, regardless of primary or secondary, are considered as primary cases, and secondary cases only include local secondary cases, i.e.,

$$\begin{aligned} N_{sec}(t-2, t+2) &= N_{sec}^{loc}(t-2, t+2) \\ N_{pri}(t-2, t+2) &= N_{pri}^{loc}(t-2, t+2) + N_{pri}^{imp}(t-2, t+2) + N_{sec}^{imp}(t-2, t+2) \\ \tilde{N}_{sec}(t-2, t+2) &= \tilde{N}_{sec}^{loc}(t-2, t+2) \end{aligned}$$

- (2) Same as (1), but local primary cases on each day is allocated to previous days according to the assumed distribution of the incubation period and contribute to the numerator for

those days. The rationale is that these local primary cases might have been infected by other cases in the previous days.

$$\begin{aligned}
N_{sec}(t-2, t+2) &= N_{sec}^{loc}(t-2, t+2) + \sum_{\tau=t+3}^T N_{pri}^{loc}(\tau, \tau) \sum_{s=t-1}^{t+2} \eta(\tau-s) \\
N_{pri}(t-2, t+2) &= N_{pri}^{loc}(t-2, t+2) + N_{pri}^{imp}(t-2, t+2) + N_{sec}^{imp}(t-2, t+2) \\
\tilde{N}_{sec}(t-2, t+2) &= \tilde{N}_{sec}^{loc}(t-2, t+2)
\end{aligned}$$

where $\eta(l)$ is the probability that the incubation period is l days (see section 1.4). We use the setting with a mean of 4 days given in Table S1.

- (3) Same as (2), but imported secondary cases will be moved from the denominator to the numerator, i.e, they are now considered as secondary cases not primary cases,

$$\begin{aligned}
N_{sec}(t-2, t+2) &= N_{sec}^{loc}(t-2, t+2) + N_{sec}^{imp}(t-2, t+2) + \sum_{\tau=t+3}^T N_{pri}^{loc}(\tau, \tau) \sum_{s=t-1}^{t+2} \eta(\tau-s) \\
N_{pri}(t-2, t+2) &= N_{pri}^{loc}(t-2, t+2) + N_{pri}^{imp}(t-2, t+2) \\
\tilde{N}_{sec}(t-2, t+2) &= \tilde{N}_{sec}^{loc}(t-2, t+2) + \tilde{N}_{sec}^{imp}(t-2, t+2)
\end{aligned}$$

It is expected that the magnitudes of the R_t for the three scenarios will have the order (1) \leq (2) \leq

(3). The confidence interval (CI) for R_t is calculated based on the CI for the Poisson mean:

$$\exp\left[\log(\widehat{R}_t) \pm 1.96 \times (N_{pri}(t-2, t+2)\widehat{R}_t)^{-1/2}\right].$$

As a sensitivity analysis, we also considered the formula $\widehat{R}_t = \frac{N_{sec}(t-2, t+2)}{N_{pri}(t-2, t+2)}$, which is the

estimation of R_t solely focusing on transmissibility of the primary cases. This estimator is more comparable to the local reproductive number estimated by the chain-binomial model in section 1.5 which also evaluates the reproductive number for a typical primary case.

1.3 The chain-binomial model

The general model Suppose we observe H transmission clusters, each with one or more primary cases with their household and/or non-household contacts, along with specific contact history between each pair of individuals. Let n_h be the size of cluster h , and let $N = \sum_{h=1}^H n_h$ be the total number of the people in these clusters. We use Λ_h to represent the collection of individuals in

cluster h . We consider two types of person-to-person close contact: frequent contact between household members and opportunistic contact between cases and non-household individuals. The probability that an infectious case infects a household member per daily contact is p_1 , and the probability that an infectious case infects a non-household individual per daily period is p_2 . In addition, each susceptible individual is subject to a constant daily infection probability of b via casual contact with the general public. Let \tilde{t}_i be the symptom onset day if an infected person i is symptomatic, and we assume \tilde{t}_i marks the peak infectivity during the infectious period. If person i is an asymptomatic infection, we also assign a \tilde{t}_i which is interpreted as the peak infectivity day. Subjects who are not infected by their last observation day T_i will have $\tilde{t}_i = \infty$. Consider the potential transmission between an infectious person j and a susceptible person i in cluster h . Let $c_{ij}(t)$ indicate whether there is a household contact (1) or a non-household contact (2) or no contact at all (0) between individuals i and j . Let $I_{\{c\}}$ be the indicator function that takes value 1 (0) if the condition c is true (false). The probability $p_{ji}(t)$ that a fully infective individual j infects a susceptible individual i on day t is determined by

$$\text{logit}(p_{ij}(t)) = I_{\{c_{ij}(t) \neq 0\}} \left[I_{\{c_{ij}(t)=1\}} \text{logit}(p_1) + I_{\{c_{ij}(t)=2\}} \text{logit}(p_2) + \boldsymbol{\beta}' \mathbf{x}_{ij}(t) \right],$$

where the logit function has the form $\text{logit}(y) = \log(y/(1-y))$, and $\mathbf{x}_{ij}(t)$ is the vector of covariates and $\boldsymbol{\beta}$ is the corresponding coefficient vector ($\boldsymbol{\beta}'$ means transpose of $\boldsymbol{\beta}$). In our notation, all bolded symbols are column vectors. Similarly, the covariate-adjusted infection probability by the casual contact with the general public is given by

$$\text{logit}(b_i(t)) = \text{logit}(b) + \boldsymbol{\alpha}' \mathbf{x}_i(t).$$

In practice, $\mathbf{x}_i(t)$ is usually a subset of $\mathbf{x}_{ij}(t)$, as the latter also encodes covariates associated with the transmission source j . For example, $\mathbf{x}_i(t)$ may be the age group and gender of individual i , which may also appear in $\mathbf{x}_{ij}(t)$. It is also common to assume coefficients in $\boldsymbol{\alpha}$ coincide with the corresponding coefficients in $\boldsymbol{\beta}$, as covariates of person i should modify his or her susceptibility in the same way regardless of the transmission source. The probability that a susceptible individual $i \in \Lambda_h$ escapes infection from all infective sources on day t is then given by $e_i(t) = (1 - b_i(t)) \sum_{j \in \Lambda_h} \phi(t - \tilde{t}_j) \theta^{1-s_j} p_{ij}(t)$, where $\phi(t - \tilde{t}_j)$ is the probability that individual j is infectious on day t given that j has symptom onset or peak infectivity on day \tilde{t}_j , which is also referred to as the relative infectivity function. Note that we assume that $\phi(l)$ depends on $l = t - \tilde{t}_j$, the distance between t and \tilde{t}_j , and that $D_{min} \leq l \leq D_{max}$, where D_{min} and D_{max} are the lower and upper bound of the infectious period in reference to the symptom onset or peak infectivity day. Note that D_{min} can be either positive, 0 or negative. If $\phi(l) > 0$ when $l < 0$, it implies infectivity during the incubation period. The symbol s_j indicates whether person j is a symptomatic case ($s_j = 1$) or an asymptomatic infection ($s_j=0$), and θ measures the relative infectivity level of an asymptomatic infection in comparison to a symptomatic case. θ can be estimated if there are a sufficient number of asymptomatic infections but is assumed known if not so. It is difficult, if not impossible, to estimate θ and $\phi(l)$ in the presence of other unknown parameters (p_1, p_2 and $\boldsymbol{\beta}$); consequently, we assume they are both known and will perform sensitivity analysis by changing their values. Denote the probability of individual i escaping infection up to day t by $Q_i(t) = \sum_{\tau=1}^t e_i(\tau)$. Suppose individual i is infected on day t . We assume the incubation period $d_i = \tilde{t}_i - t$ has a known discrete distribution $\eta(l) = Pr(d_i = l)$, $d_{min} \leq l \leq d_{max}$, where d_{min} and d_{max} are the minimum and maximum duration of the incubation period. If individual i is an asymptomatic infection, δ_i cannot be called incubation

period, but we assume δ_i follows the same distribution. The distribution of the incubation period will be derived from literature. Define $\tilde{\mathbf{t}}_h = \{\tilde{t}_i: i \in \Lambda_h\}$ as the collection of symptom onset days (or peak infectivity days for asymptomatic infections). If $\tilde{\mathbf{t}}_h$ is fully observed, we can construct the likelihood contributed by person i as

$$L_{(i)}(b, p_1, p_2, \boldsymbol{\beta} | \tilde{\mathbf{t}}_h) = \begin{cases} Q_i(T_i), & \text{if not infected,} \\ \sum_{t=\tilde{t}_i-d_{max}}^{\tilde{t}_i-d_{min}} \{\eta(t - (\tilde{t}_i - d_{max})) [1 - e_i(t)] Q_i(t - 1)\}, & \text{if infected.} \end{cases}$$

It is important to note that the design of our study is called case-ascertained design, which implies that each cluster is observed because of the primary case in the cluster. For this type of design, the infection of primary cases will not contribute to the likelihood, but their infectious periods contributed to the risk of infection of their cluster members and thus contributed to the likelihoods of those individuals. For simplicity, in each cluster, day 1 in the likelihood corresponds to the actual day $\tilde{t}_{\Lambda_h} - d_{max} + 1$, where \tilde{t}_{Λ_h} is the earliest symptom onset day among primary cases of the cluster. For proper inference, one has to condition the likelihood on the fact that all cluster members who are not primary cases have not had symptom onset by day \tilde{t}_{Λ_h} . Under this condition, day $\tilde{t}_{\Lambda_h} - d_{max} + 1$ is the first day with uncertainty about the infection status of a cluster member who is not the primary case (Yang 2006).

The E-M algorithm For asymptomatic infections, we do not observe symptom onset. According to our assumption, there exists a peak infectivity day \tilde{t}_i and that the time lag between \tilde{t}_i and the infection time t also follows the known distribution of the incubation period. However, we do not observe \tilde{t}_i , and we use an Expectation-Maximization (EM) algorithm to integrate out such uncertainty (Tan1992). Briefly, let \mathbf{U}_h be the collection of \tilde{t}_i for all individuals $i \in \Lambda_h$ and who

are asymptomatic infections. In this analysis, we consider two possibilities settings for the range of possible \tilde{t}_i for each asymptomatic infection; one is from $\tilde{t}_{\Lambda_h} - d_{max} + 1$ to T_i , and the other is from $\tilde{t}_{\Lambda_h} - d_{min}$ to T_i . Let \mathbf{O}_h be the collection of \tilde{t}_i for all individuals $i \in \Lambda_h$ and who are symptomatic cases. \mathbf{U}_h and \mathbf{O}_h represent the unobserved and observed outcomes in cluster h . Let \mathbf{u}_{hl} , $l = 1, \dots, \delta_h$ be all possible realizations of \mathbf{U}_h , where δ_h is the number of such realizations. Let $\boldsymbol{\psi} = (b, p_1, p_2, \boldsymbol{\beta})$ denote all the parameters. Rewrite the individual likelihood as $L_{(i)}(\boldsymbol{\psi}|\mathbf{O}_h, \mathbf{U}_h)$, and define the cluster-level and population-level likelihood based on complete data as $L_h(\boldsymbol{\psi}|\mathbf{O}_h, \mathbf{U}_h) = \prod_{i \in \Lambda_h} L_{(i)}(\boldsymbol{\psi}|\mathbf{O}_h, \mathbf{U}_h)$ and $L(\boldsymbol{\psi}|\mathbf{O}, \mathbf{U}) = \prod_{h=1}^H L_h(\boldsymbol{\psi}|\mathbf{O}_h, \mathbf{U}_h)$, respectively, where $\mathbf{O} = \{\mathbf{O}_h, h = 1, \dots, H\}$ and $\mathbf{U} = \{\mathbf{U}_h, h = 1, \dots, H\}$ represent population-level observed and missing outcomes. The EM algorithm proceeds as follows.

1. Choose an initial value $\boldsymbol{\psi}^{(0)}$, and set $\hat{\boldsymbol{\psi}}^{(0)} = \boldsymbol{\psi}^{(0)}$.
2. At iteration $r \geq 1$, update the conditional probabilities $\lambda_{hk}^{(r)} = \frac{L_h(\hat{\boldsymbol{\psi}}^{(r)}|\mathbf{O}_h, \mathbf{u}_{hk})}{\sum_{l=1}^{\delta_h} L_h(\hat{\boldsymbol{\psi}}^{(r)}|\mathbf{O}_h, \mathbf{u}_{hl})}$, $k = 1, \dots, \delta_h$. For clusters with completely observed outcomes, we have $\delta_h = 1$ and $\lambda_{h1}^{(r)} = 1$.
3. Maximize $\Omega(\boldsymbol{\psi}, \hat{\boldsymbol{\psi}}^{(r)}) = \sum_{h=1}^H \sum_{k=1}^{\delta_h} \lambda_{hk}^{(r)} \ln L_h(\boldsymbol{\psi}|\mathbf{O}_h, \mathbf{u}_{hk})$ with regard to $\boldsymbol{\psi}$ to find $\hat{\boldsymbol{\psi}}^{(r+1)}$.
4. Repeat 2 and 3 until convergence in the estimates $\hat{\boldsymbol{\psi}}^{(r)}$ of $\boldsymbol{\psi}$.

Direct calculation of the covariance matrix is difficult as one need to sum over all possible realizations of the missing data for the whole study population, which is not a linear operation for the calculation of the missing information (Louis1982). Our solution is to estimate the covariance using a sampling approach (yang2014). Specifically, we sample K sets of missing data \mathbf{U}_h for each cluster h from the distribution $Pr(\mathbf{U}_h|\mathbf{O}_h, \hat{\boldsymbol{\psi}})$, where K is a large integer (e.g.,

1000) and $\hat{\boldsymbol{\psi}}$ is the final parameter estimate. Let these samples be denoted by $\hat{\boldsymbol{u}}_{hk}$, $h = 1, \dots, H$, $k = 1, \dots, K$. The covariance matrix, $\hat{\boldsymbol{V}}$, is given by

$$\hat{\boldsymbol{V}}^{-1} = -\frac{1}{K} \sum_{k=1}^K \frac{d^2}{d\boldsymbol{\psi}^2} \ln L(\boldsymbol{\psi} | \boldsymbol{O}, \hat{\boldsymbol{U}}_{.k}) - \left[\frac{1}{K} \sum_{k=1}^K \left(\frac{d}{d\boldsymbol{\psi}} \ln L(\boldsymbol{\psi} | \boldsymbol{O}, \hat{\boldsymbol{U}}_{.k}) \right) \left(\frac{d}{d\boldsymbol{\psi}} \ln L(\boldsymbol{\psi} | \boldsymbol{O}, \hat{\boldsymbol{U}}_{.k}) \right)' - \left(\frac{1}{K} \sum_{k=1}^K \frac{d}{d\boldsymbol{\psi}} \ln L(\boldsymbol{\psi} | \boldsymbol{O}, \hat{\boldsymbol{U}}_{.k}) \right) \left(\frac{1}{K} \sum_{k=1}^K \frac{d}{d\boldsymbol{\psi}} \ln L(\boldsymbol{\psi} | \boldsymbol{O}, \hat{\boldsymbol{U}}_{.k}) \right)' \right],$$

Where $\hat{\boldsymbol{U}}_{.k} = \{\hat{\boldsymbol{u}}_{hk} : h = 1, \dots, H\}$, and $\ln L(\boldsymbol{\psi} | \boldsymbol{O}, \hat{\boldsymbol{U}}_{.k}) = \sum_{h=1}^H \ln L(\boldsymbol{\psi} | \boldsymbol{O}, \hat{\boldsymbol{u}}_{hk})$.

In our cluster data, there are only six asymptomatic infections, and the number of asymptomatic infection in each cluster is at most two, which makes the EM algorithm possible. If there are more asymptomatic infections such that enumeration of all possible realizations in each cluster is impossible, one can use the Monte-Carlo EM (MCEM) algorithm.

1.4 The natural history of disease

The natural history of disease is represented by the density function of the incubation period, $\eta(l)$, and the relative infectivity function $\phi(l)$. The incubation period has been estimated to have a mean of 4-7 days with a wide range. Although there have been unofficially reported incubation periods as long as more than 20 days, these are likely rare events, and most countries adopted a two-week quarantine policy. For these reasons, we assume a maximum incubation period of 14 days and a minimum duration of 1 days. Based on our unpublished parametric Weibull models for the incubation period for contact tracing clusters of cases in China (before), we generated four possible incubation period settings with means of 4, 5, 6 and 7 days and standard deviations of 2.0, 2.5, 3.0 and 3.5 days (Table S1).

Much is unknown about the infectious period of COVID-19. The median time from symptom onset to hospitalization before Jan. 26, 2009 was about 4 days based on surveillance data

provided by China CDC, suggesting a relatively short exposure period of contacts to cases after their symptom onset. We chose three settings for the relative infectivity profiles with maximum durations of 13, 16 and 19 days (including 5 days before symptom onset) for the infectious period (Table S2). When the relative infectivity levels during the illness period are nonincreasing, they can be thought of as the upper-tail cumulative distribution (also called survival function) of a random infectious period (starting from the symptom onset day), and the corresponding mean durations of such as random post-onset infectious period are 4.1, 6.9 and 8.2 days. These relative infectivity levels do not imply biological infectiousness of a host; rather, they reflect the combined effect of biological infectiousness and contact restrictions due to, e.g., hospitalization or social distancing. It was suspected that an infected person is actually infectious during the incubation period. This is partially supported by our unpublished findings that the mean serial interval and generation interval are shorter than the mean incubation period. To account for this possibility, we assume an infected person can be infectious as early as 5 days before symptom onset. Although the relative infectivity level is set to 1.0 for both pre-onset days and a few post-onset days, the actual infectivity levels will differ. We introduce a time-dependent binary covariate for each infected person, and this indicator takes value 0 for the incubation period and 1 for the illness period starting from the onset day. This incubation indicator variable is assumed to affect infectivity such that the actual infectivity will differ between the incubation period and the illness period, with the difference measured by the coefficient for the indicator. The relative infectivity level during the incubation period of a symptomatic case is further adjusted for the uncertainty in the length of the incubation period. Specifically, let $\phi(l)$ be the nominal relative infectivity level assigned to the incubation period, where $l = -5, -4, \dots, -1$ is the day in reference to the symptom onset day, i.e., 5, 4 and up to 1 day before symptom onset.

The effective relative infectivity on day l is the product of $\phi(l)$ and the probability that the incubation period is longer than or equal to $|l|$ days, i.e., $\phi(l) \sum_{j=|l|}^{d_{max}} \eta(j)$. For asymptomatic infection, we make a simple assumption that the relative infectivity level is identical to the nominal levels during the whole infectious period of a symptomatic case, but without adjusting for uncertainty in the incubation period or the time-dependent indicator for illness. This assumption, though likely unrealistic, should have limited impact on the estimates of SARs and covariate effects because of the scarceness of asymptomatic infections in our data.

1.5 SAR and local reproductive number

We report SAR and local reproductive number estimates for the model adjusting for no covariates except for the time-dependent incubation indicator which affects infectivity of an infected person. For each day l , let $\phi^*(l) = \phi(l) \left[\sum_{j=|l|}^{d_{max}} \eta(j) \right]^{I_{\{l < 0\}}}$ be the effective relative infectivity function, and let $p_k^*(l) = (p_k)^{I_{\{l < 0\}}} \left(\frac{p_k OR}{1 - p_k + p_k OR} \right)^{I_{\{l \geq 0\}}}$ be the effective transmission probability. SAR is calculated as

$$SAR_k = 1 - \prod_{l=D_{min}}^{D_{max}} [1 - p_k^*(l) \phi^*(l)], k = 1, 2,$$

where OR is the odds ratio for the incubation indicator. The covariate-adjusted transmission probability $\frac{p_k OR}{1 - p_k + p_k OR}$ is derived from the logistic regression $\text{logit}^{-1}[\text{logit}(p_k) + \log(OR)]$, where $\text{logit}(p_k) = \log[p/(1 - p)]$ and logit^{-1} is the inverse logit transformation. With the incubation indicator, p_k is interpreted as the average daily transmission probability from an infective person to a susceptible person ($k = 1$ for household and 2 for non-household) during the incubation period of the infective, and $\frac{p_k OR}{1 - p_k + p_k OR}$ is the daily transmission probability when the infective is ill and completely infectious ($\phi(l) = 1$). The SAR among non-household

contacts, SAR_2 , may be a less appropriate measure for transmissibility than the daily transmission probabilities p_2 or $\frac{p_2OR}{1-p_2+p_2OR}$, as some types of non-household contact do not last over the whole infectious period by nature and could even be one-time event, e.g., contact with other restaurant customers or flight passengers. For this reason, we present both p_k and $\frac{p_kOR}{1-p_k+p_kOR}$ in Table S3.

The local reproductive number is defined as the mean number of infections a case can generate during his or her entire infectious period via both close household and non-household contact. This reproductive number can be viewed as an approximate to the basic reproductive number R_0 if the whole population is susceptible and no intervention is implemented. Due to the tight control of human movement in Guangzhou during the study period, this local reproductive number does not reflect R_0 in our study. The distributions of daily numbers of household and non-household contacts of cases are shown in figure S2. Let $n_1(l)$ and $n_2(l)$ be the average numbers of household and non-household contacts per primary case on day l , respectively, with the symptom onset day of the primary case as day 0:

Day	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13
$n_1(l)$	3.58	3.56	3.49	3.46	3.43	3.24	2.68	2.29	1.98	1.45	1.09	0.91	0.63	0.46	0.35	0.2	0.19	0.11	0.1
$n_2(l)$	2.0	1.95	1.97	1.87	1.8	2.3	1.05	0.92	0.79	0.48	0.31	0.26	0.22	0.16	0.16	0.15	0.04	0.04	0.015

We stop at day 13 as it is the maximum value of D_{max} in our settings for the infectious period.

The local reproductive number is calculated as

$$R = \sum_{k=1}^2 \sum_{l=D_{min}}^{D_{max}} \{n_k(l)p_k^*(l)\phi^*(l) \prod_{m=D_{min}}^{l-1} [1 - p_k^*(l)\phi^*(l)]\},$$

where $n_k(l) \prod_{m=D_{min}}^{l-1} [1 - p_k^*(l)\phi^*(l)]$ is the expected number of susceptible contacts by day $l - 1$ among the $n_k(l)$ contacts made on day l . As the actual observed $n_k(l)$'s are used, this R represents the local reproductive number under the implemented control measures. In particular, the $n_k(l)$'s decreased quickly after day 0 (symptom onset day of the primary case), indicating the effect of isolation or quarantine of both identified cases and their close contacts. To estimate local reproductive numbers in the absence of quarantine (movement constraint may remain) had there been no quarantine, we recalculate R by projecting the average contact numbers during days -5 to 0 to the subsequent days 1 to 13:

Day	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	
$n_1(l)$	3.58	3.56	3.49	3.46	3.43	3.24	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46	3.46
$n_2(l)$	2.0	1.95	1.97	1.87	1.8	2.3	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99	1.99

1.6 Assessment of goodness-of-fit

We had developed a goodness-of-fit measure for chain-binomial models, which compares the observed and model-fitted frequencies of symptom onsets among exposed person-days (yang2006). The calculation of model-fitted, or expected, frequency of symptom onset for each exposed person-day is conditioning on transmission dynamics that have already realized up to the day before. In the presence of asymptomatic infections, this approach is likely inadequate. Instead, we propose to compare observed and expected frequencies of infection, not symptom onset, for each exposed person-day. The expected frequency for a person-day (i, t) is simply $[1 - e_i(t)]Q_i(t - 1)$. The observed frequency is obtained by allocating each observed symptom onset to the possible infection days, i.e., from $\tilde{t}_i - d_{max}$ to $\tilde{t}_i - d_{min}$, using certain weights.

More specifically, the weight for a given $l = \tilde{t}_i - t$, where $t \in [\tilde{t}_i - d_{max}, \tilde{t}_i - d_{min}]$, is given by $w_l = \phi(l)[1 - e_i(t)] \prod_{\tau=\tilde{t}_i-d_{max}}^{t-1} e_i(\tau)$, and the weights are then normalized to have sum 1. These weights for observed frequencies depend on model parameter estimates via $e_i(t)$ in order to account for the fact that the number of infectious individuals on each day should contribute to the allocation of a symptomatic case to the potential infection days of this case. For asymptomatic infections, we choose the most likely peak infectivity day \tilde{t}_i corresponding to value of \mathbf{u}_{hl} associated with the highest likelihood at the cluster level, and the calculation of weights proceeds as for symptomatic infections. The observed and expected frequencies of infections are then aggregated over the whole study population by day, where all clusters are aligned with $\tilde{t}_{\Lambda_h} - d_{max} + 1$ as day 1 for each cluster. The 95% confidence intervals (CI) may be calculated for the daily aggregated expected frequencies, but such 95% CIs tend to be overly narrow due to the reduced uncertainty as a result of conditioning the calculation for each exposed person-day on observed transmission dynamics before that day. We adopted a frequently used alternative for constructing marginal, not conditional, 95% CIs by simulating transmissions within clusters using the model-fitted parameters. We plot the observed and expected daily frequencies of infections together with the pointwise 2.5% and 97.5% quantiles of daily numbers of infections in the simulated outbreaks, as a visual diagnosis for the goodness-of-fit of our models.

Table S1. Probability densities for the distribution of the incubation period.

Mean	Days from infection to symptom onset or peak infectivity													
Duration														
(days)	1	2	3	4	5	6	7	8	9	10	11	12	13	14
4	0.091	0.16	0.19	0.18	0.15	0.10	0.061	0.032	0.015	0.0061	0.0022	7.2×10^{-4}	2.1×10^{-4}	5.3×10^{-5}
5	0.058	0.11	0.14	0.16	0.15	0.13	0.10	0.068	0.044	0.026	0.014	0.0072	0.0034	0.0015
6	0.043	0.079	0.11	0.12	0.13	0.12	0.11	0.088	0.070	0.052	0.037	0.025	0.016	0.0098
7	0.040	0.065	0.082	0.093	0.098	0.098	0.095	0.088	0.080	0.071	0.061	0.052	0.043	0.035

Table S2. Relative infectivity levels during the infectious period.

Max	Days from symptom onset or peak infectivity day											
Duration												
(Days)	-5 ~ 2	3	4	5	6	7	8	9	10	11	12	13
13	1.0	0.8	0.6	0.4	0.2	0.1						
16	1.0	1.0	1.0	0.8	0.8	0.6	0.4	0.2	0.1			
19	1.0	1.0	1.0	1.0	0.8	0.8	0.6	0.4	0.2	0.2	0.1	0.05

Table S3. Model-based estimates (and 95% confidence intervals) of secondary attack rates (SAR) among household and non-household contacts, and model-based estimates of local reproductive number (local R) with and without quarantine. Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease. This model is not adjusted for age group.

Mean Incubation Period (days)	Max Infectious Period (days)	Household defined by close relatives				Household defined by residential address			
		SAR (%)		Local Reproductive Number		SAR (%)		Local Reproductive Number	
		Household	Non-household	With quarantine	Without quarantine	Household	Non-household	With quarantine	Without quarantine
4	13	13.8 (11.1-17.0)	7.1 (4.7-10.6)	0.48 (0.39-0.58)	0.62 (0.51-0.75)	19.3 (15.5-23.9)	5.3 (3.6-7.8)	0.43 (0.35-0.53)	0.54 (0.44-0.67)
	16	15.8 (12.5-19.7)	8.3 (5.4-12.6)	0.48 (0.39-0.60)	0.71 (0.57-0.88)	22.1 (17.5-27.5)	6.2 (4.1-9.1)	0.43 (0.35-0.54)	0.62 (0.50-0.78)
	19	16.8 (13.3-21.1)	9.0 (5.8-13.6)	0.48 (0.39-0.60)	0.76 (0.61-0.95)	23.4 (18.4-29.3)	6.6 (4.4-9.8)	0.44 (0.35-0.55)	0.66 (0.52-0.84)
5	13	12.9 (10.3-16.0)	6.6 (4.3-9.8)	0.48 (0.39-0.59)	0.58 (0.47-0.71)	18.0 (14.3-22.3)	4.9 (3.3-7.3)	0.43 (0.35-0.53)	0.50 (0.41-0.62)
	16	14.3 (11.2-18.0)	7.4 (4.8-11.2)	0.48 (0.38-0.61)	0.64 (0.51-0.80)	19.8 (15.5-24.9)	5.5 (3.6-8.2)	0.43 (0.34-0.54)	0.56 (0.44-0.70)
	19	14.9 (11.6-18.9)	7.8 (5.0-11.9)	0.48 (0.38-0.61)	0.67 (0.53-0.85)	20.7 (16.0-26.2)	5.7 (3.8-8.6)	0.43 (0.34-0.55)	0.58 (0.45-0.74)
6	13	12.1 (9.7-15.1)	6.2 (4.1-9.4)	0.48 (0.39-0.59)	0.54 (0.44-0.67)	16.9 (13.5-21.1)	4.7 (3.2-6.9)	0.43 (0.34-0.53)	0.48 (0.38-0.59)
	16	13.1 (10.3-16.5)	6.8 (4.4-10.4)	0.48 (0.38-0.61)	0.59 (0.47-0.74)	18.2 (14.2-23.0)	5.1 (3.4-7.6)	0.43 (0.34-0.54)	0.51 (0.40-0.65)
	19	13.5 (10.5-17.3)	7.1 (4.5-10.8)	0.48 (0.38-0.62)	0.61 (0.48-0.77)	18.7 (14.5-23.9)	5.3 (3.5- 7.9)	0.43 (0.34-0.55)	0.53 (0.41-0.68)
7	13	11.7 (9.4-14.5)	6.2 (4.1-9.2)	0.48 (0.39-0.59)	0.53 (0.43-0.65)	16.4 (13.1-20.3)	4.7 (3.2-6.8)	0.43 (0.35-0.53)	0.46 (0.37-0.57)
	16	12.5 (9.8-15.7)	6.6 (4.3-10.0)	0.49 (0.39-0.61)	0.56 (0.45-0.71)	17.4 (13.6-21.9)	5.0 (3.4-7.3)	0.43 (0.34-0.54)	0.49 (0.39-0.62)
	19	12.8 (9.9-16.3)	6.8 (4.4-10.4)	0.49 (0.38-0.62)	0.58 (0.45-0.74)	17.8 (13.8-22.6)	5.1 (3.4-7.6)	0.43 (0.34-0.55)	0.51 (0.40-0.64)

Table S4. Model-based estimates (and 95% confidence intervals) of daily transmission probabilities for household contacts (p_1^*) and non-household contacts (p_2^*) during the incubation and illness periods. Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease. This model is not adjusted for age group.

Incubation Period (days)	Max Infectious Period (days)	Household contact defined by close relatives				Household contact defined by residential address			
		Incubation Period		Illness Period		Incubation Period		Illness Period	
		p_1^*	p_2^*	p_1^*	p_2^*	p_1^*	p_2^*	p_1^*	p_2^*
4	13	1.58 (1.10-2.25)	0.79 (0.49-1.27)	1.78 (1.19-2.66)	0.89 (0.51-1.55)	2.40 (1.70-3.39)	0.62 (0.39-0.96)	2.49 (1.63-3.79)	0.64 (0.37-1.11)
	16	1.67 (1.19-2.34)	0.85 (0.54-1.34)	1.41 (0.93-2.12)	0.71 (0.40-1.26)	2.53 (1.83-3.51)	0.65 (0.42-1.00)	1.99 (1.29-3.05)	0.51 (0.29-0.89)
	19	1.69 (1.21-2.35)	0.87 (0.55-1.36)	1.34 (0.89-2.03)	0.69 (0.39-1.22)	2.56 (1.85-3.53)	0.66 (0.43-1.01)	1.89 (1.23-2.90)	0.49 (0.28-0.85)
5	13	1.86 (1.37-2.52)	0.92 (0.59-1.42)	1.23 (0.70-2.16)	0.61 (0.30-1.21)	2.77 (2.05-3.72)	0.71 (0.47-1.08)	1.68 (0.92-3.05)	0.43 (0.21-0.87)
	16	1.93 (1.45-2.57)	0.96 (0.63-1.47)	0.96 (0.55-1.68)	0.48 (0.24-0.96)	2.86 (2.16-3.79)	0.74 (0.50-1.10)	1.32 (0.73-2.38)	0.34 (0.17-0.68)
	19	1.95 (1.47-2.58)	0.98 (0.64-1.49)	0.91 (0.52-1.59)	0.45 (0.22-0.91)	2.88 (2.18-3.81)	0.75 (0.50-1.11)	1.25 (0.69-2.25)	0.32 (0.16-0.64)
6	13	2.06 (1.57-2.69)	1.03 (0.68-1.57)	0.80 (0.36-1.80)	0.40 (0.16-1.00)	3.01 (2.31-3.92)	0.80 (0.54-1.17)	1.08 (0.46-2.53)	0.28 (0.11-0.72)
	16	2.09 (1.62-2.70)	1.06 (0.70-1.59)	0.64 (0.30-1.36)	0.32 (0.13-0.77)	3.06 (2.37-3.94)	0.81 (0.55-1.18)	0.87 (0.39-1.92)	0.23 (0.09-0.55)
	19	2.10 (1.63-2.71)	1.07 (0.71-1.60)	0.60 (0.28-1.28)	0.30 (0.12-0.73)	3.07 (2.39-3.95)	0.82 (0.56-1.19)	0.82 (0.37-1.81)	0.21 (0.09-0.51)
7	13	2.16 (1.68-2.77)	1.11 (0.73-1.67)	0.56 (0.20-1.60)	0.29 (0.09-0.89)	3.13 (2.45-4.01)	0.85 (0.58-1.23)	0.77 (0.26-2.22)	0.20 (0.07-0.64)
	16	2.18 (1.71-2.77)	1.12 (0.75-1.68)	0.46 (0.18-1.19)	0.24 (0.08-0.68)	3.15 (2.48-4.00)	0.85 (0.59-1.24)	0.63 (0.24-1.66)	0.17 (0.06-0.48)
	19	2.18 (1.72-2.77)	1.13 (0.76-1.69)	0.44 (0.17-1.12)	0.22 (0.08-0.64)	3.16 (2.49-4.01)	0.86 (0.59-1.24)	0.60 (0.23-1.56)	0.16 (0.06-0.45)

Table S5. Model-based estimates (and 95% confidence intervals) of daily probability of infection from an external source (b) and the odds ratios for the relative infectivity during the illness versus incubation period. Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease. This model is not adjusted for age group.

Mean Incubation Period (days)	Max Infectious Period (days)	Household contact defined by close relatives		Household contact defined by residential address	
		b ($\times 10^{-4}$)	Odds Ratio	b ($\times 10^{-4}$)	Odds Ratio
4	13	2.23 (1.14, 4.34)	1.13 (0.59, 2.18)	2.24 (1.14, 4.37)	1.04 (0.54, 2.00)
	16	2.02 (0.99, 4.12)	0.84 (0.44, 1.59)	2.02 (0.99, 4.12)	0.78 (0.41, 1.48)
	19	1.95 (0.95, 4.01)	0.79 (0.42, 1.49)	1.96 (0.96, 4.03)	0.73 (0.39, 1.39)
5	13	1.76 (0.81, 3.87)	0.66 (0.31, 1.41)	1.81 (0.82, 3.95)	0.60 (0.27, 1.32)
	16	1.59 (0.70, 3.62)	0.49 (0.23, 1.03)	1.64 (0.73, 3.71)	0.45 (0.21, 0.97)
	19	1.56 (0.68, 3.57)	0.46 (0.22, 0.96)	1.62 (0.72, 3.67)	0.42 (0.20, 0.90)
6	13	1.55 (0.64, 3.74)	0.39 (0.15, 1.02)	1.56 (0.64, 3.81)	0.35 (0.13, 0.97)
	16	1.42 (0.57, 3.51)	0.30 (0.12, 0.74)	1.44 (0.58, 3.58)	0.28 (0.11, 0.70)
	19	1.40 (0.57, 3.48)	0.28 (0.11, 0.69)	1.43 (0.57, 3.56)	0.26 (0.10, 0.66)
7	13	1.53 (0.61, 3.86)	0.26 (0.08, 0.84)	1.52 (0.59, 3.88)	0.24 (0.07, 0.79)
	16	1.41 (0.54, 3.63)	0.21 (0.07, 0.61)	1.40 (0.54, 3.65)	0.20 (0.07, 0.57)
	19	1.40 (0.54, 3.60)	0.20 (0.07, 0.57)	1.39 (0.54, 3.63)	0.18 (0.06, 0.54)

Table S6. Model-based odds ratios (and 95% confidence intervals) for age effects on susceptibility and infectivity and relative infectivity during the illness period in comparison to the incubation period. Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease.

Incubation Period (days)	Max Infectious Period (days)	Household contact defined by close relatives				Household contact defined by residential address			
		Susceptibility		Infectivity		Susceptibility		Infectivity	
		<20 vs. ≥60	20-59 vs. ≥60	<60 vs. ≥60	Illness vs. Incubation	<20 vs. ≥60	<20 vs. ≥60	<60 vs. ≥60	Illness vs. Incubation
4	13	0.27 (0.13-0.55)	0.80 (0.53-1.19)	0.79 (0.51-1.23)	1.14 (0.59-2.21)	0.23 (0.11-0.47)	0.74 (0.49-1.10)	0.97 (0.61-1.54)	1.03 (0.53-1.98)
	16	0.27 (0.13-0.54)	0.79 (0.53-1.18)	0.79 (0.52-1.23)	0.84 (0.44-1.59)	0.23 (0.11-0.47)	0.74 (0.49-1.10)	0.97 (0.62-1.52)	0.77 (0.40-1.45)
	19	0.27 (0.13-0.54)	0.79 (0.53-1.18)	0.79 (0.52-1.22)	0.79 (0.42-1.49)	0.23 (0.11-0.47)	0.74 (0.49-1.10)	0.97 (0.62-1.52)	0.72 (0.38-1.36)
5	13	0.27 (0.13-0.54)	0.80 (0.54-1.20)	0.76 (0.50-1.18)	0.66 (0.31-1.42)	0.23 (0.11-0.46)	0.74 (0.50-1.11)	0.91 (0.58-1.43)	0.60 (0.27-1.31)
	16	0.26 (0.13-0.54)	0.80 (0.54-1.20)	0.76 (0.50-1.17)	0.49 (0.23-1.02)	0.23 (0.11-0.46)	0.74 (0.50-1.11)	0.91 (0.59-1.42)	0.44 (0.21-0.95)
	19	0.26 (0.13-0.54)	0.80 (0.54-1.20)	0.76 (0.50-1.17)	0.45 (0.22-0.95)	0.23 (0.11-0.46)	0.74 (0.50-1.11)	0.91 (0.59-1.42)	0.41 (0.19-0.88)
6	13	0.26 (0.13-0.54)	0.81 (0.54-1.21)	0.74 (0.48-1.13)	0.39 (0.15-1.02)	0.23 (0.11-0.46)	0.75 (0.50-1.12)	0.88 (0.57-1.37)	0.35 (0.13-0.96)
	16	0.26 (0.13-0.54)	0.81 (0.54-1.20)	0.74 (0.48-1.13)	0.30 (0.12-0.73)	0.23 (0.11-0.46)	0.75 (0.50-1.12)	0.88 (0.57-1.36)	0.27 (0.11-0.69)
	19	0.26 (0.13-0.54)	0.81 (0.54-1.20)	0.74 (0.48-1.13)	0.28 (0.11-0.68)	0.23 (0.11-0.46)	0.75 (0.50-1.12)	0.88 (0.57-1.36)	0.25 (0.10-0.64)
7	13	0.26 (0.13-0.54)	0.81 (0.54-1.21)	0.73 (0.48-1.12)	0.27 (0.085-0.82)	0.22 (0.11-0.46)	0.75 (0.50-1.12)	0.87 (0.56-1.35)	0.24 (0.076-0.77)
	16	0.26 (0.13-0.54)	0.81 (0.54-1.21)	0.73 (0.48-1.12)	0.21 (0.073-0.59)	0.22 (0.11-0.46)	0.75 (0.50-1.12)	0.87 (0.56-1.34)	0.19 (0.067-0.56)
	19	0.26 (0.13-0.54)	0.81 (0.54-1.21)	0.73 (0.48-1.11)	0.19 (0.069-0.55)	0.22 (0.11-0.46)	0.75 (0.50-1.12)	0.87 (0.56-1.34)	0.18 (0.063-0.52)

Table S7. Model-based estimates (and 95% confidence intervals) of daily probability of infection from an external source (b) and daily transmission probabilities during the incubation period of a COVID-19 case for household contacts (p_1) and non-household contacts (p_2). Estimates are reported using two different definitions of household contact (close relatives, or only individuals sharing the same residential address) and for all investigated settings of the natural history of disease. This model is adjusted for age group, and thus the probabilities are for susceptible elderly ≥ 60 years old when they are exposed to infective elderly ≥ 60 years old.

Mean Incubation Period (days)	Max Infectious Period (days)	Household contact defined by close relatives			Household contact defined by residential address		
		$b (\times 10^{-4})$	$p_1 (\times 10^{-2})$	$p_2 (\times 10^{-2})$	$b (\times 10^{-4})$	$p_1 (\times 10^{-2})$	$p_2 (\times 10^{-2})$
4	13	2.83 (1.35, 5.95)	2.59 (1.56, 4.27)	1.14 (0.62, 2.11)	3.04 (1.45, 6.39)	3.76 (2.27, 6.17)	0.86 (0.47, 1.58)
	16	2.56 (1.17, 5.60)	2.76 (1.69, 4.50)	1.24 (0.68, 2.26)	2.73 (1.25, 5.97)	3.97 (2.42, 6.45)	0.92 (0.51, 1.67)
	19	2.48 (1.13, 5.47)	2.79 (1.71, 4.54)	1.27 (0.70, 2.30)	2.67 (1.22, 5.85)	4.01 (2.45, 6.51)	0.93 (0.51, 1.69)
5	13	2.21 (0.94, 5.18)	3.12 (1.95, 4.96)	1.37 (0.76, 2.46)	2.42 (1.04, 5.66)	4.46 (2.78, 7.07)	1.04 (0.58, 1.86)
	16	2.00 (0.83, 4.85)	3.26 (2.06, 5.13)	1.45 (0.81, 2.58)	2.21 (0.92, 5.30)	4.63 (2.91, 7.27)	1.09 (0.61, 1.92)
	19	1.98 (0.82, 4.79)	3.29 (2.08, 5.16)	1.47 (0.82, 2.61)	2.19 (0.91, 5.24)	4.66 (2.94, 7.32)	1.09 (0.62, 1.94)
6	13	1.91 (0.74, 4.94)	3.51 (2.25, 5.44)	1.57 (0.88, 2.78)	2.06 (0.79, 5.37)	4.94 (3.16, 7.66)	1.19 (0.68, 2.09)
	16	1.77 (0.67, 4.65)	3.59 (2.32, 5.53)	1.62 (0.92, 2.85)	1.91 (0.72, 5.05)	5.03 (3.23, 7.76)	1.22 (0.70, 2.12)
	19	1.76 (0.67, 4.62)	3.61 (2.33, 5.56)	1.64 (0.93, 2.87)	1.90 (0.72, 5.02)	5.05 (3.25, 7.79)	1.22 (0.70, 2.13)
7	13	1.87 (0.69, 5.02)	3.71 (2.41, 5.68)	1.71 (0.97, 3.00)	1.99 (0.73, 5.40)	5.17 (3.34, 7.92)	1.28 (0.73, 2.22)
	16	1.73 (0.63, 4.74)	3.76 (2.45, 5.74)	1.74 (0.99, 3.05)	1.85 (0.67, 5.09)	5.23 (3.39, 7.98)	1.30 (0.75, 2.24)
	19	1.73 (0.63, 4.72)	3.78 (2.46, 5.75)	1.75 (1.00, 3.06)	1.84 (0.67, 5.07)	5.24 (3.40, 7.99)	1.30 (0.75, 2.25)

Figure S1. Spatial distribution of COVID-19 clusters with at least one non-primary cases. For each cluster, the outer ring indicates whether the primary case was imported (red) or local (blue) and whether the symptom onset of the primary case was on/before (dark color) or after (light color) Jan. 23, 2020, the day when lockdown of Wuhan was initiated. The proportions of household (green) and non-household (purple) cases are shown in the inner circle for each cluster. Township-level population densities are shown as the background.

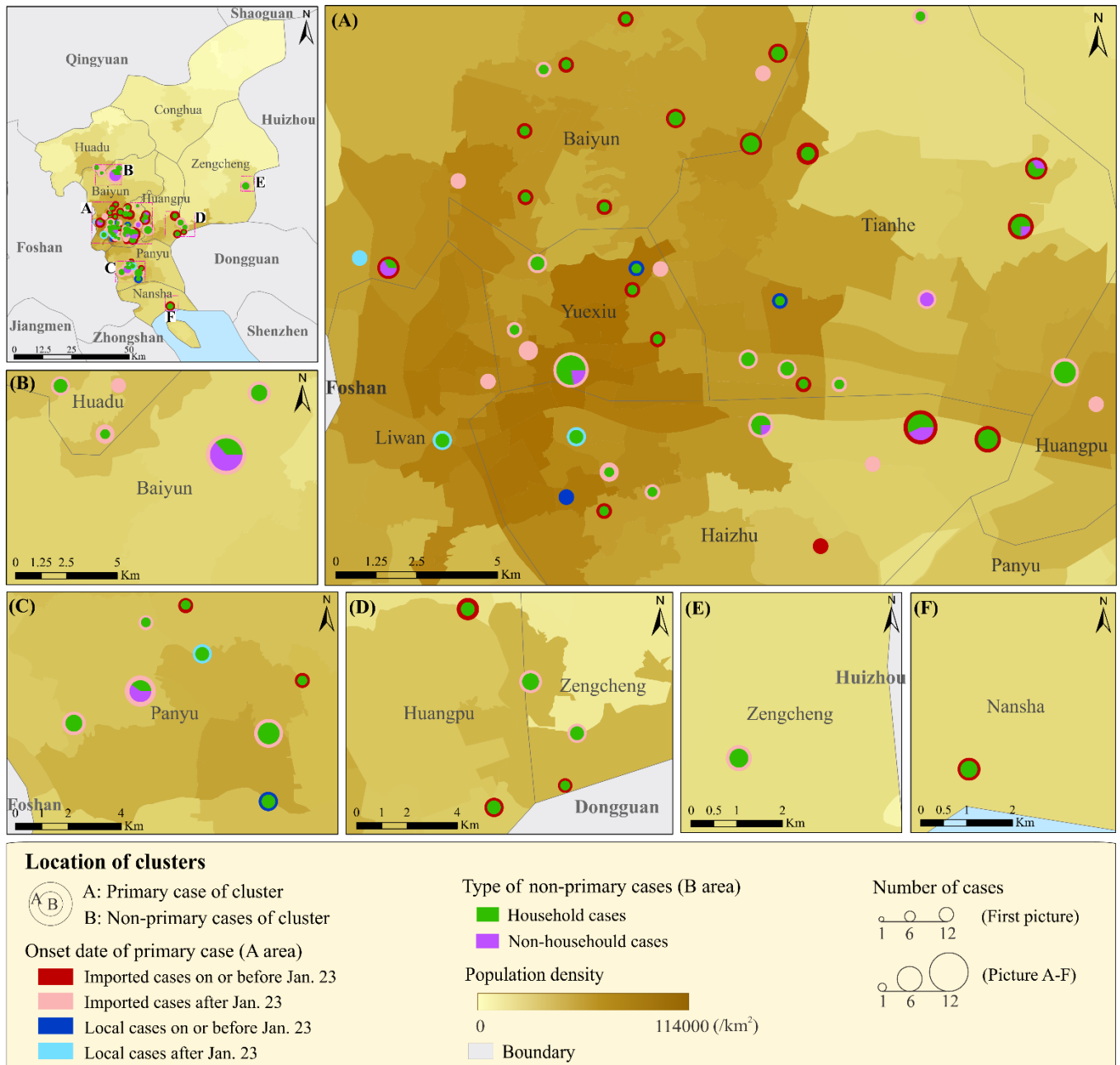


Figure S2. Distribution of number of household (upper) and non-household (lower, log-scale) contacts across all clusters in Guangzhou, China over the potential infectious period of the primary case with the symptom onset day of the primary case set as day 0. For clusters with more than 1 primary cases, the numbers of contacts are averaged over the co-primary cases.

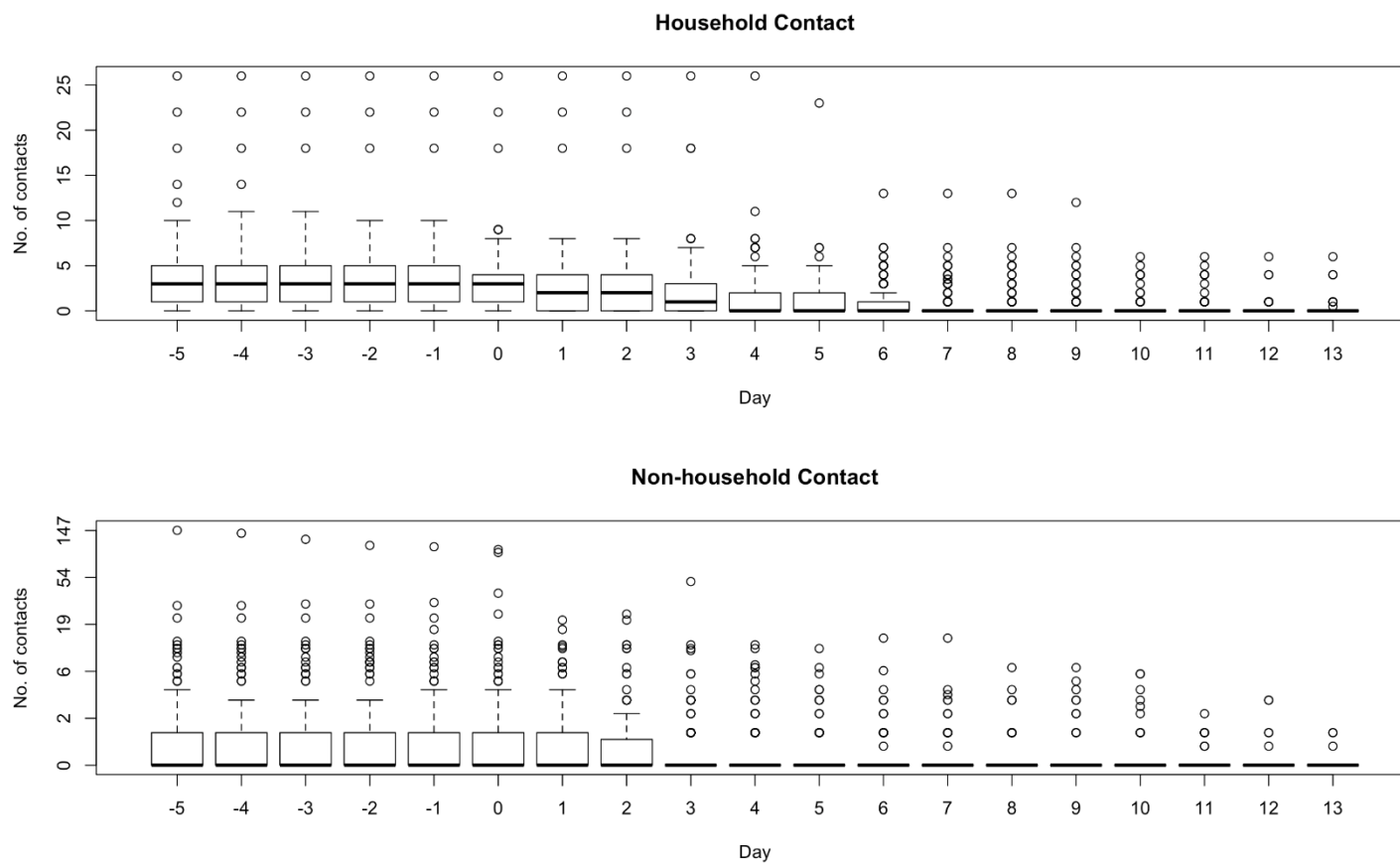


Figure S3. Observed (red triangles) and model-fitted (red solid curve) numbers of infections, together with the 95% CIs for the model-fitted numbers (blue dashed curves), on each day for the whole study population for different settings of the incubation and infectious periods. All clusters were aligned in time by the symptom onset day of the earliest primary case and day 0 is 13 days before that day. SSD stands for sum of squares of differences between observed and model-fitted numbers. The observed daily infection numbers changed slightly by the settings of the incubation and infectious periods because the way we allocate cases to their possible infection days also depend on the settings and model parameter estimates (see Appendix 1.6)

